

DOCUMENT RESUME

ED 340 720

TM 017 690

AUTHOR Glowacki, Margaret L.
TITLE An Analysis of Test Equating Models for the Alabama High School Graduation Examination.
PUB DATE Nov 91
NOTE 27p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (20th, Lexington, KY, November 12-15, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Comparative Testing; Cutting Scores; *Equated Scores; *Graduation Requirements; High Schools; High School Students; *Item Response Theory; Language Tests; Mathematical Models; Mathematics Tests; Reading Tests
IDENTIFIERS *Alabama High School Graduation Examination; Angoff Methods; Equipercentile Equating; Linear Equating Method; Score Distribution; Smoothing Methods; Test Equivalence

ABSTRACT

The purpose of this study was to determine which equating models are appropriate for the Alabama High School Graduation Examination (AHSGE) by equating two previously administered fall forms for each subject area of the AHSGE and determining whether differences exist in the test score distributions or passing scores resulting from the equating models. The two forms equated were Form 10 and Form 30, two of the original preequated forms of the AHSGE administered at fall administrations one year apart. The equating models examined were: (1) linear; (2) equipercentile; and (3) one-parameter, two-parameter, and three-parameter item response theory (IRT) models. The equating design used two random samples of 10,000 students each from non-equivalent populations and a set of common items. Any of the five models appear to be appropriate for equating forms of the AHSGE reading, mathematics, and language subject area tests developed in the same manner as the two forms that were equated, and using the same preequating methodology. It would seem most appropriate to use: (1) conventional methods for equating since the test forms were developed using conventional methods; and (2) IRT models only if they offered something that non-IRT models did not. The importance of smoothing is discussed. Three graphs and one table present study data, and a 43-item list of references is included. (SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED340720

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

MARGARET L. GLOWACKI

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

AN ANALYSIS OF TEST EQUATING MODELS FOR THE
ALABAMA HIGH SCHOOL GRADUATION EXAMINATION

by

Margaret L. Glowacki
Evaluation and Assessment Laboratory
The University of Alabama

Paper presented at the annual meeting of the Mid-South Educational Research Association, Lexington,
Kentucky, November 13-15, 1991.

TM017690

An Analysis of Test Equating Models for the Alabama High School Graduation Examination

Introduction

In the last two decades many factors have contributed to the development and refinement of new designs and models for equating test scores. Among these factors are the increase in the number and variety of testing programs that use interchangeable forms of a test, the need for security and fairness when administering multiple forms of a test, the development of item response theory and its applications to test equating, test disclosure requirements in some states, and the need to address issues raised by testing critics (Brennan, 1987; Gilmer, 1987; Holland & Rubin, 1982; Marco, 1981).

Test equating is the process of adjusting the raw score scale of one test form to correspond to the raw score scale of another form of the same test and establish equivalent scores. There are two major types of test equating, horizontal and vertical. Horizontal equating is the process of establishing equivalent scores on two or more forms of the same test at the same ability level. Vertical equating is the process of determining equivalent scores on two or more forms of a test measuring the same attribute but at different educational levels (Holmes, 1982). This study focused on horizontal equating.

Angoff (1971) discussed two restrictions that apply to test equating. First, tests being equated should be measures of the same characteristic or ability. Second, the conversion of the score distributions should be independent of the examinees used to collect the data for the conversion and should be applicable to all situations. Lord (1980) discussed four conditions he believed should be met if scores on two tests were to be equated: same ability, equity, population invariance, and symmetry. There is general agreement among researchers that the population invariance and symmetry conditions need to be satisfied, but disagreement exists regarding whether the equity and same-ability conditions need to be met (Petersen, Kolen, & Hoover, 1989).

Marco (1981) maintained that tests could be equated successfully only if the tests being equated measured the same underlying knowledge, skills, or abilities. If the tests met this condition, then the tests were assumed to yield perfectly related scores except for measurement error. Lord (1982) demonstrated statistically that generally it is not possible for two tests to be perfectly equated.

Multiple forms of a test generally are constructed to be as similar as possible in content, format, item type, and difficulty (Petersen, Cook, & Stocking, 1983; Petersen, Kolen, & Hoover, 1989; Skaggs & Lissitz, 1986). Gulliksen (1950) maintained that approximately parallel test forms could be developed by plotting item difficulties and item discrimination indices, identifying items clustered together on the plot, and randomly assigning one item in each cluster to each form. The test forms also should be parallel in item content.

Jaeger (1981) maintained that "regardless of the care with which they are constructed, multiple test forms invariably differ in difficulty level and in variability" (p. 23). Marco (1981) questioned the effectiveness of methods of constructing parallel forms of a test. He examined test scores from five forms of the Graduate Record Examination Aptitude Test administered between January, 1980 and February, 1981. These forms were created using fairly rigorous statistical specifications. Marco noted that a raw score of 70 converted to a scaled score of 770 on one form and 810 on another. Without equating these tests, it would not be fair to the examinee taking the more difficult form.

This is why equating is so important to test takers--it ensures, as nearly as possible, that the reported test scores are fair expressions of underlying knowledge, abilities, or skills regardless of the test form actually used. (Marco, 1981, p. 108)

Angoff (1971) stated that

since two forms of a test can rarely if ever be made to be precisely equivalent in level and range of difficulty, it becomes necessary to **equate** the forms . . . so that scores derived from the two forms **after conversion** will be directly equivalent. (p. 562)

Marco (1981) defined an equating method as

an empirical procedure for determining a transformation to be applied to scores on a test form to make them equivalent to scores on another test form. Because an

equating method is an empirical procedure, it involves a design for data collection and statistical procedures or models for determining the transformation. (p. 113)

Data collection designs include single-group, counter-balanced random-groups, equivalent groups, anchor-test--random-groups, anchor-test--nonequivalent-groups, section pre-equating, and item pre-equating (Petersen, Kolen, and Hoover, 1989). Statistical models include equipercentile equating, linear equating, one-, two-, and three-parameter item response theory (IRT) models.

Comparison Studies

Many studies have been conducted to compare equating models under a variety of conditions (Cook & Eignor, 1983; Eignor, Stocking, & Cook, 1990; Hills, Subhiyah, & Hirsch, 1988; Kolen, 1981; Lawrence & Dorans, 1990; Livingston, Dorans, & Wright, 1990; Lord & Wingersky, 1984; Marco, Petersen, & Stewart, 1983; Modu, 1982; Petersen, Cook, & Stocking, 1983; Schmitt, Cook, Dorans, & Eignor, 1990; Skaggs & Lissitz, 1985). The equating models commonly examined included equipercentile, linear, and one-, two-, and three-parameter IRT models. Results of the comparative studies were mixed.

The majority of studies examined test equating models for equating norm-referenced tests. Only one study examined test equating models for equating minimum competency tests. Hills, Subhiyah, and Hirsch (1988) compared linear equating models and the one- and three-parameter equating models using scores from the SSAT-II, a minimum competency test used in the Florida Statewide Student Assessment Testing program. The 1986 test was equated to the 1984 test. The study attempted to determine how well the models being examined equated test scores that had highly negatively skewed test score distributions. Results indicated that, despite extremely skewed distributions, all models except the three-parameter model using the fixed b's calibration method gave similar results. The authors concluded that for minimum competency tests not developed using IRT methods, linear equating models were appropriate, and for those programs that use IRT for other areas of the testing program, IRT equating models could be adopted.

Several researchers found that equipercentile, linear, and IRT equating models worked equally well, while others concluded that IRT models were superior, linear models produced acceptable results, and equipercentile equating models were inadequate. Many factors appeared to affect how well each equating model performed including parameter estimation problems, differences in data collection designs, different linking processes, multidimensionality of test forms, test length, test content, and the distribution of item difficulty and sample ability parameters. Most of the studies examined equating models under different conditions, making it difficult to compare the results of the studies. Ongoing research is examining how these conditions affect equating in general, rather than comparing the equating models. Few issues regarding test equating have been adequately resolved. Further research appears warranted in many areas, including minimum competency tests.

Purpose of the Study

The purpose of this study was to determine which equating models are appropriate for the Alabama High School Graduation Examination (AHSGE) by equating two previously administered fall forms for each subject area of the AHSGE and determining if differences exist in the test score distributions or passing scores resulting from the equating models. The two forms equated were Form 10 and Form 30, two of the original pre-equated forms of the AHSGE administered at fall administrations one year apart. The equating models examined were linear, equipercentile, and one-, two-, and three-parameter item response theory models. The equating design used random samples from non-equivalent populations and a set of common items.

Research Questions

1) Will linear equating, equipercentile equating, and one-, two-, and three-parameter item response theory equating models produce different raw test score distributions or differences at the passing score when equating two pre-equated fall forms of the AHSGE reading test?

2) Will linear equating, equipercentile equating, and one-, two-, and three-parameter item response theory equating models produce different raw test score distributions or differences at the passing score when equating two pre-equated fall forms of the AHSGE mathematics test?

3) Will linear equating, equipercentile equating, and one-, two-, and three-parameter item response theory equating models produce different raw test score distributions or differences at the passing score when equating two pre-equated fall forms of the AHSGE language test?

Method

Instrument

The data used in this study were from administrations of the Alabama High School Graduation Examination (AHSGE). The AHSGE is an untimed, objectives-based minimum competency examination consisting of three subject area tests: reading, mathematics, and language. The skills in these three areas have been defined as minimum skills that all students should master before completing high school. Each subject area test is scored independently and students are required to pass all three tests to receive an Alabama high school diploma. Once a student has taken the AHSGE, he or she is required to retake only those subject area tests not previously passed (Alabama State Department of Education, 1988).

In an attempt to ensure fairness and security of the examination from administration to administration, a parallel form of the AHSGE is administered each fall and spring. Students are allowed to take the examination when they become juniors and are provided four opportunities while in school to pass the AHSGE. If all sections of the AHSGE are not passed by the time a student's class graduates, the student does not receive a diploma. The student may return to the high school to take any failed sections of the examination during subsequent administrations.

Each fall, all juniors and any students who previously have failed or have not taken sections of the examination are administered a parallel form of the AHSGE. Each spring, a different form of the

test is administered to juniors who did not pass one or more sections during the fall administration and any students who previously have failed or not taken sections of the examination. The spring populations generally are smaller than the fall populations and mainly consist of students who have failed one or more sections of the AHSGE, which means that a difference in ability levels between the spring and fall populations may exist (Lockwood, 1985). Lockwood maintained that use of an equating procedure "that adjusts item difficulties based on differential performance between two similar groups of examinees is not appropriate in this case" (p. 27).

Due to the differences in abilities of the fall and spring populations, the first four forms of the AHSGE were pre-equated. "In this context test pre-equating merely refers to a system of test development that assures identical item formats and content and virtually identical item difficulties between test forms" (Lockwood and McLean, 1984, p. 2). Two forms were developed for each subject area "that were meticulously matched in terms of item content and format and which were virtually identical in item difficulties (the sum of difficulties within .1)" (Lockwood and McLean, 1984, p. 3). The two forms for each subject area consisted of equivalent halves. The first form was comprised of two halves, A1 and A2, and the second form was comprised of two halves, B1 and B2. A1 was pre-equated to be equivalent to B2, and A2 was pre-equated to be equivalent to B1. The two forms for each subject area were pilot tested and then used to create four parallel forms for each subject area. The four forms were Form 10, Form 20, Form 30 and Form 40. Form 10 consisted of halves A1 and A2, Form 20 consisted of halves A1 and B1, Form 30 consisted of halves A2 and B2, and Form 40 consisted of halves B1 and B2. Form 10 was administered during Fall 1983 and Form 30 was administered during Fall 1984.

Cronbach's (1970) coefficient alpha was used to estimate the internal consistency or reliability of Forms 10 and 30 for all subject areas. The coefficient alphas for Form 10 of the reading, mathematics, and language tests were, respectively, .973, .964, and .972. The coefficient alphas for

Form 30 of the reading, mathematics, and language tests were, respectively, .944, .950, and .945. The coefficient alphas indicate a high degree of internal consistency for forms 10 and 30 in all three subject areas. Three aspects of validity were established: content, curricular, and instructional validity (Alabama State Board of Education, 1988; Hardy, 1984; McLean & Morse, 1985).

Subjects

The data used were from students classified as Alabama public high school juniors or above when they took Forms 10 and 30 of the AHSGE as part of the requirements for obtaining an Alabama public high school diploma. Two random samples of 10,000 students each were selected, one from each population of approximately 50,000 students for the two test administrations. Prior to selecting the random samples, all examinees who had not attempted one or more subtests for the first time during the Fall 1983 or Fall 1984 administrations were removed from the populations. A SAS (SAS Institute Inc., 1985) program for selecting a random sample (Johnson, 1990) was used to select the random samples. The same samples were used for all three subject area tests.

Table 1 presents descriptive statistics for the samples selected from Form 10 and Form 30 for the reading, mathematics, and language tests. The means and standard deviations are very similar for the reading and mathematics tests, but differ for the language test. This difference resulted from the difference in the length of the 1983 language test and the 1984 language test. The test was shortened from 119 items in 1983 to 100 items in 1984. The passing score was adjusted from 92 in 1983 to 77 in 1984.

Table 1.

Sample Descriptive Statistics for Fall 1983 and Fall 1984 Administrations of the Alabama High School Graduation Examination

Subject Area	Form	N	Number of Items	Mean	SD	Correlation of Common Items
Reading	10	10000	80	74.01	9.966	0.96
	30	10000	80	74.95	8.179	0.95
Mathematics	10	10000	95	77.65	15.344	0.98
	30	10000	95	79.12	14.056	0.97
Language	10	10000	119	102.34	16.906	0.97
	30	10000	100	87.97	12.299	0.96

Data Collection

The data collection design used in this study was the nonequivalent-populations--common-items design as described by Angoff (1971). This design uses two randomly sampled groups from different populations. A minimum of 20 items or 20% of the number of items on the test, whichever is larger, is recommended for the set of common items. The AHSGE reading test forms have 31 common items, the AHSGE mathematics test forms have 40 common items, and the AHSGE language test forms have 41 common items.

Techniques Compared

Linear Equating: Linear equating is based on the assumption that two forms of a test constructed to be parallel will have nearly identical raw score distributions except for possible differences in the means and variances of the distributions (Crocker & Algina, 1986). If all moments of the score distribution beyond these first two are equal, then the distributions will have the same shape. If this assumption is true, equivalent scores can be identified by determining pairs of scores, one score from Form 10 and one score from Form 30 that have the same standard or z-score. The only requirement of linear equating is the calculation of a linear function to transform scores from one scale to another

using some form of the equation $Y = AX + B$, where A is the slope of the conversion line, and B is the intercept or the point where the conversion line intersects the Y axis.

The basic linear method for nonequivalent groups not widely different in ability from Design IV as described by Angoff (1971) was used for the linear equating. This method is based on the assumptions of univariate selection theory. A SAS (SAS Institute Inc., 1985) program was written to perform the linear equating for reading, language, and math.

Equipercentile Equating

Equipercentile equating is a curvilinear model in which the raw scores on two tests are considered to be equated if they correspond to the same percentile rank for a group of examinees (Angoff, 1971). A transformation is performed to convert the sets of scores. When the raw score distributions for two tests are identical in shape, a table can be constructed showing pairs of raw scores with identical percentile ranks. The pairs of raw scores will not be numerically equal, so transformation of one set of raw scores to correspond to the other set to produce equated scores is necessary. Both sets of raw scores also can be converted to a different set of scores (Petersen, Kolen, & Hoover, 1989).

No assumptions are made about the test forms being equated when equipercentile equating is viewed as a mathematical model. Scores on one test form are stretched or compressed so that its raw score distribution coincides with the raw score distribution of the other test form (Petersen, Kolen, & Hoover, 1989). Equipercentile equating is not as straightforward as it may seem and is susceptible to sampling error. The equipercentile model is data dependent and when sample sizes are small, it is possible to have zero frequencies for some scores within the observed score range, and only a few observations for the extreme scores (Cook & Petersen, 1987). Even with large sample sizes when the raw score conversion is more regular, finding two raw scores with identical relative cumulative frequencies is difficult, because seldom will there be a raw score on one form of a test having the

same percentile rank as a raw score on the second form of the test. The raw score distributions must be made continuous to find pairs of raw scores with the same percentile rank. This is accomplished by the subjective selection of interpolation or smoothing methods (Cook & Petersen, 1987).

Two types of smoothing methods are available, presmoothing and postsmoothing. Presmoothing techniques include moving medians (Tukey, 1977), rolling weighted averages (Angoff, 1971; Cureton & Tukey, 1951), and the negative hypergeometric distribution (Keats & Lord, 1962). Postsampling techniques include rolling weighted averages (Angoff, 1971; Cureton & Tukey, 1951), cubic splines, and regression-based polynomial functions (Reinsch, 1967). Fairbank (1987) examined fourteen techniques for smoothing. He found negative hypergeometric presmoothing to be the most effective presmoothing technique and orthogonal regression and cubic splines to be effective postsamplers.

The curvilinear method for groups not widely different in ability in Angoff's (1971) Design IV was used for the equipercentile equating. This method is based on the same assumptions of univariate selection theory as the linear method. The equipercentile equating was performed using programs written in SAS (SAS Institute Inc., 1985). Many of the scores on Form 10 did not have equivalent scores on Form 30, so analytical presmoothing using the three-point moving weighted averages method was performed to create a full range of scores (Tukey, 1977).

Item Response Theory Equating

Item response theory (IRT), also known as latent trait theory, is based on the tenet that there are latent or unobserved traits possessed by individuals that account for the way they respond to test items. IRT models specify a relationship between a latent trait and an individual's observed performance on a test designed to measure that latent trait. This relationship is described by a mathematical function called an item characteristic curve (ICC) (Crocker & Algina, 1986). The ICC "plots the probability of responding correctly to an item as a function of the latent trait (denoted by θ) underlying performance on the items on a test" (Crocker & Algina, 1986, p. 340).

Several IRT models have been developed, but the one-, two-, and three-parameter logistic models have been most commonly used in testing applications (Skaggs & Lissitz, 1986). All of these models use ability estimates and differ primarily in the parameters believed to affect performance. In the one-parameter model, also known as the Rasch model, the probability of a correct response is dependent upon item difficulty. In the two-parameter model the probability of a correct response is dependent upon item difficulty and item discrimination. The Rasch model is often considered a special case of the two-parameter model. The three-parameter model uses a pseudo-guessing parameter in addition to the item difficulty and discrimination parameters (Crocker & Algina, 1986).

The computer program LOGIST 5 (Wingersky, Barton, & Lord, 1982) was used to conduct the one-, two-, and three-parameter item response theory (IRT) equating. LOGIST 5 produces estimates of difficulty (b), discrimination (a), and guessing (c) parameters for each item and an ability estimate (θ) for each examinee. Theta (θ) and b are placed on scales with a mean of 0 and a standard deviation of 1. The estimates are obtained using a maximum likelihood procedure with special procedures for handling omitted items (Lord, 1974).

For all three IRT models, the concurrent calibration method was used to estimate the item parameters. Concurrent calibration requires a group of common examinees or common items, and parameters for the two tests being equated are estimated simultaneously. This results in item parameter estimates and ability estimates being placed on a common scale (Petersen, Kolen, & Hoover, 1989). For the one-parameter model, only b was estimated while a and c were held fixed. For the two-parameter model, a and b were estimated while c was held fixed. For the three-parameter model, all three item parameters were estimated.

For the three IRT models, when ability levels were estimated for the examinees, ability levels were not estimated for any examinees with perfect scores or answering less than approximately one-third of the items on the test. LOGIST offers the choice of including all examinees, but accurate

ability estimates cannot be calculated for examinees with perfect scores or where there is little information about an examinee's ability.

The LOGIST printout provided the parameters for each item, and ability levels were provided for all examinees who did not receive a perfect score and who answered more than one-third of the test items. In cases where traditional means of reporting scores are preferred over the reporting of ability estimates, it is possible to transform the ability estimates to estimated number-right true scores, estimated true score formula scores, or estimated number-right observed score frequency distributions. The ideal transformation would be to estimate observed formula score frequency distributions, but this cannot be obtained using IRT methods (Cook & Eignor, 1981).

Each of the available methods has problems associated with it. The disadvantage of using true score estimates is that when using the three-parameter IRT model, the possible range of scores only extends at the lower level of the distribution to the chance level. This would be a problem for testing programs reporting scores below the chance level, although it does not present a problem for the AHSGE, as examinees either pass or fail according to the set passing score. The disadvantage of the estimated number-right observed score method is that the conversions obtained may not be applicable to observed formula scores.

It has not been determined whether observed score equating or true score equating produces better results. Kolen (1981) conducted a cross-validation study in which the results of the two procedures varied. Lord and Wingersky (1984) conducted a study in which the two procedures gave very similar results. For the purposes of this study, the estimated number-right true score method was used to translate θ to estimated true scores.

Comparison of the Models

The best criterion against which to judge equating results is not known. However, several measures have been derived. These include equating a test form to itself, comparing the results of the

equating to a conventional method if one is known to function well (Marco, Petersen, & Stewart, 1983), or examining the stability rather than the accuracy of equating (Kolen, 1981). Since no equating methods have previously been tested with the AHSGE, and very few studies have been conducted concerning equating of minimum competency tests, the equating results were compared to the actual raw score distribution on Form 10.

The equating model results were compared by examining the means, standard deviations, variances, and equated passing scores. The Form 10 raw score distribution and the distributions resulting from the various equating models were graphed and compared.

Results

For reading, the means and standard deviations for the scaled scores differed little from the criterion (Form 10) mean and standard deviation. The five equating models produced means and standard deviations less than the criterion. Differences between the criterion and the scaled score means ranged from -0.01 for the equipercentile model to -0.17 for the two-parameter IRT model. The differences between the passing score on the criterion (Form 10) and the equated passing scores ranged from -0.09 for the one-parameter IRT model to 0.40 for the linear and equipercentile models.

For mathematics, the means and standard deviations for the scaled scores differed little from the mean and standard deviation for the criterion (Form 10) scores. The five equating models produced means greater than the criterion mean, and standard deviations less than the criterion standard deviation. Differences between the criterion and the scaled score means ranged from 0.31 for the two-parameter IRT model to 0.42 for the linear and equipercentile models. The differences between the passing score on Form 10 and the equated passing scores ranged from 0.42 for the three-parameter IRT model to 0.70 for the equipercentile model.

For language, the means and standard deviations for the scaled scores differed from the mean and standard deviation for the criterion scores (Form 10) because of the difference in the number of

items on the two tests. Form 10 had 119 items, and Form 30 had 100 items. The five equating models produced means less than the criterion mean, which would be expected due to the smaller number of items. There was little difference between the scaled score means. The scaled score means ranged from 86.05 for the equipercentile and one-parameter IRT models to 86.32 for the three-parameter IRT model, a difference of 0.27. The passing score on Form 10 was 92, and this was adjusted to 77 on Form 30. The equated passing scores for the five methods were within 0.72 points of 77. The passing score of 77.42 for the one-parameter IRT model was closest to the passing score of 77, and the passing score for the linear model 77.72 was the greatest distance from 77.

For each raw score on Form 10 there is a final scaled score resulting from each of the equating models. For the reading tests, the largest differences between the raw and scaled scores were found with the equipercentile model at the lower end of the distribution where less than 1% of the examinees scored. Above the 5th percentile, all five models gave results within 1 point of each other. For the math tests, the largest differences between the raw and scaled scores were found with the equipercentile model at the lower end of the distribution where less than 1% of the examinees scored. Above the 7th percentile, all five methods gave results within 1 point of each other. For the language tests, the largest differences between the raw and scaled scores were found with the equipercentile model at the lower end of the distribution where less than 1% of the examinees scored. Above the 5th percentile, all five models give results within 1 point of each other.

Graphs of the criterion and scaled score distributions for the three subject area tests are presented in Figures 1-3. The graphs show that for reading and mathematics, the final conversions of the five equating models are very similar to the criterion distribution. For language, the graph shows the final conversions of the five equating models differ from the criterion distribution, which would be expected due to the different number of items, but are very similar to each other. For all subject areas, the largest discrepancies appear at the lower end of the distribution for the equipercentile

model. There is no distribution for the three-parameter model below the chance scoring level. For the purpose of reporting and plotting the score distributions, if a true score estimate was not provided for a raw score on Form 10 from the IRT ability levels, the true score estimates were interpolated from the estimated true score distributions.

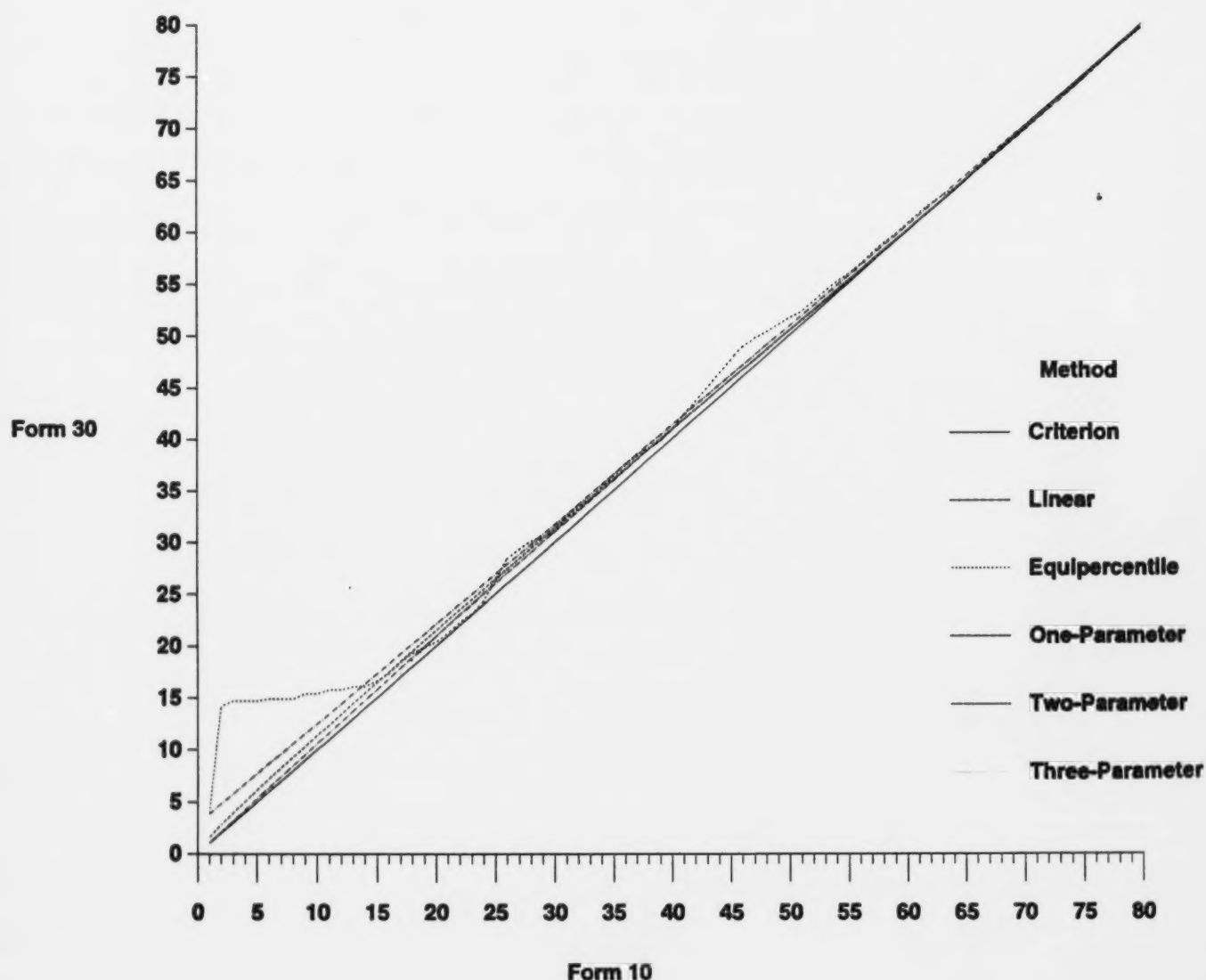


Figure 1. Plot of equated test score distributions for the AHSGE reading test.

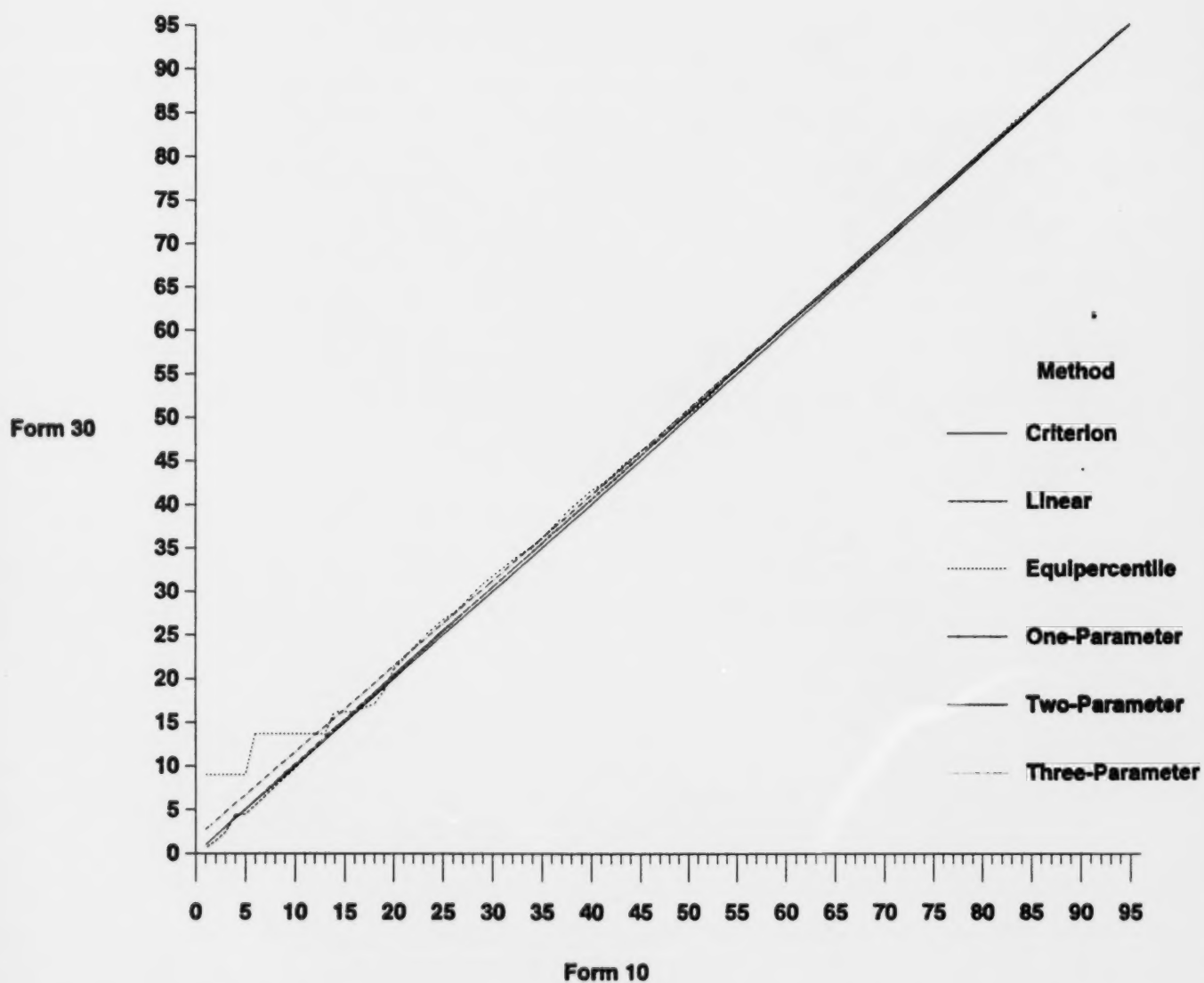


Figure 2. Plot of equated test score distributions for the AHSGE mathematics test.

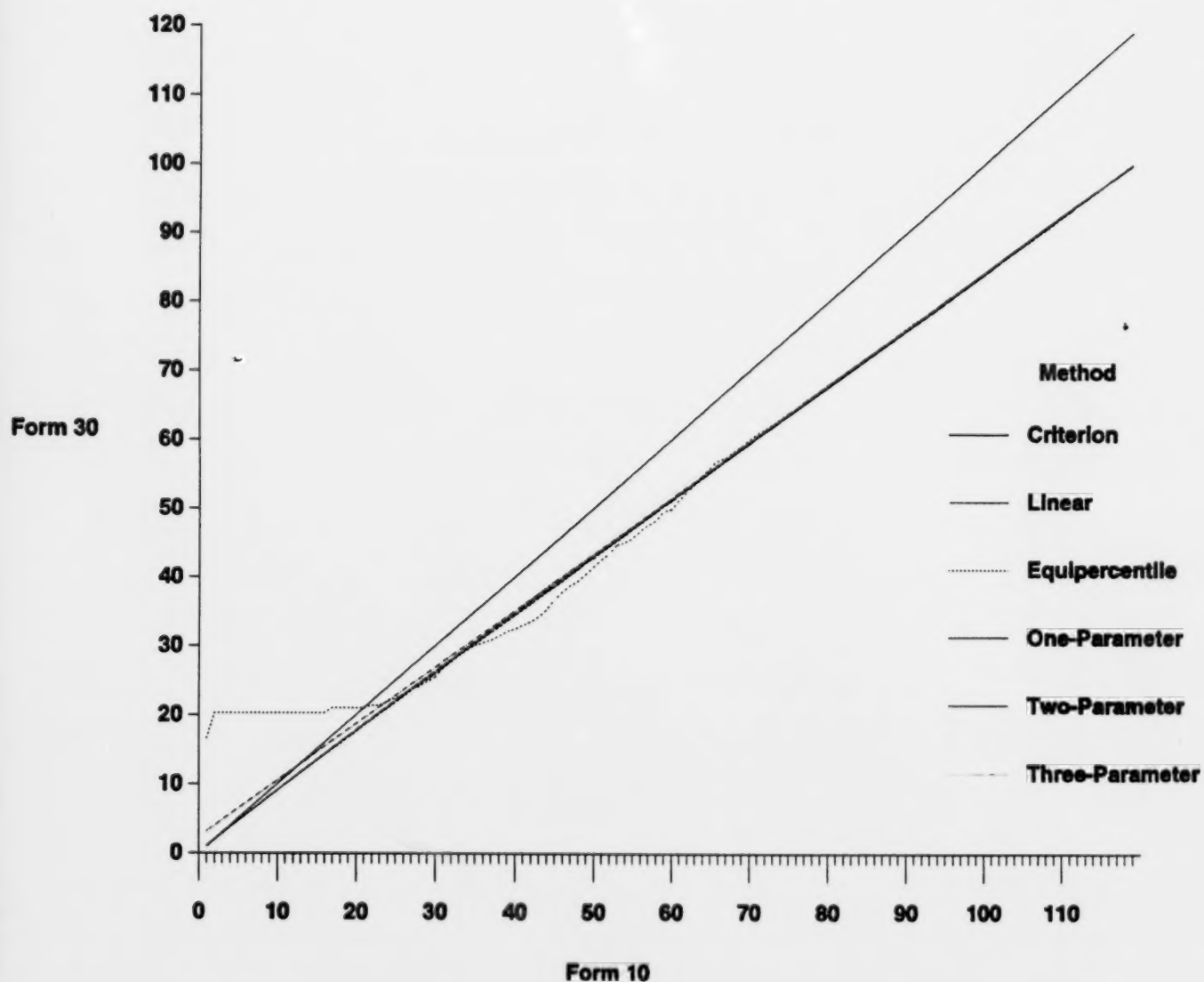


Figure 3. Plot of equated test score distributions for the AHSGE language test.

Conclusions and Discussion

The major conclusion of this study is that any of the five equating models examined appears to be appropriate for equating forms of the AHSGE reading, mathematics, and language subject area tests developed in the same manner as the two forms that were equated, and using the same pre-equating methodology. All of the models produced similar results, although the results varied at different points throughout the test score distributions for each model. The one-parameter IRT model produced scores closest to the criterion for both the reading and mathematics tests at the lower end of the distribution. The closest scores produced throughout the rest of the distribution varied among the models. Although it was not possible to determine which model produced the closest scores for the language test due to the difference in the number of items, the distributions resulting from the equatings were very similar.

The equipercentile model appeared to perform as well as the other equating models around the passing score and at the higher end of the distributions, but less adequately at the lower end of the score distribution for all of the subject areas. This is due to the small number of observations occurring at the lower end of the score distributions.

The three-parameter IRT model provided good results at the passing score, resulting in the most closely approximated passing score for reading and mathematics, and at the higher levels of the distributions, although true score estimates below the chance scoring level were not provided. The results of this study were not affected by the performance of the equipercentile or three-parameter IRT equating models at the lower end of the score distributions. It is the passing scores, rather than the scores at the lower end of the distributions, that are of interest for the AHSGE since examinees either pass or fail according to the passing score.

The passing score is of major interest when equating the AHSGE. As stated previously, all models produced good results for all three subject areas at the passing score, although the

three-parameter IRT model provided an exact approximation of the passing score for the reading test and the closest approximation for the mathematics test. The one-parameter IRT model provided the closest approximation for the language test.

Several factors may account for the close results of the five equating models. Minimum competency tests such as the AHSGE generally can be expected to produce similar test score distributions. The tests are designed to measure minimum skills, and it is expected that the majority of examinees have achieved these skills. This results in a reduction of the variability of the scores.

The pre-equating of the AHSGE test forms resulted in tests that theoretically have met the assumptions of classical true score theory for parallel tests. The content of the tests was divided into competencies, and within each competency, items with nearly identical p-values and similar content were placed on each form. This resulted in forms that were nearly identical in item content and difficulty. The discrimination indexes for the selected items also were very similar. This may have affected the results of the three IRT equating models, which were very close. Although examinees are encouraged to guess if they are unsure of an item on the AHSGE, the results of the three IRT models indicated that the guessing and discrimination parameters of the two- and three-parameter IRT models had little impact on the equating results.

Although, the results of this study indicated that any of the five equating models examined would be appropriate for equating forms of the AHSGE, it would be most appropriate to use conventional methods for equating since the test forms were developed using conventional methods. Cook and Eignor (1981) stated that when conventional procedures perform adequately for a test that has been developed using conventional methods, there is no reason not to use conventional methods for equating the test unless they prove to be unsatisfactory. It would seem reasonable to use IRT models only if they offered something non-IRT equating models did not. It may be appropriate to go to the

expense of an IRT equating and the possibility of non-fitting items, if it can be justified that the IRT equating can serve to check the conventional equating results.

Both the linear and equipercentile models showed similar results in this study, but the linear model is the preferred model. Angoff (1971) contended that linear equating is the model to be preferred if it can be assumed that the differences in the shapes of the raw score distributions are small enough to be disregarded. Due to the nature of minimum competency testing and the pre-equating of the AHSGE, it can be assumed that any differences in the shapes of the raw score distributions of the tests are small enough to be disregarded.

According to Jaeger (1981), the linear model produces more precise results than the equipercentile model at the extremes of a score distribution that is unimodal and reasonably symmetrical. This was found to be true in this study. The linear model produced closer score approximations at the lower end of the test score distributions. The equipercentile model most likely produced results similar to the linear model at the middle and upper levels of the distributions because of the analytical smoothing that was performed. If the smoothing had not been performed, or another smoothing technique had been selected, the equipercentile equating results may have been different.

The purpose of the smoothing was to create a continuous score distribution so all scores on Form 10 would have equivalent scores on Form 30. If the equipercentile model were chosen as the model to be used for equating in a minimum competency testing program such as the AHSGE, smoothing methods should be examined more closely.

When deciding on an equating model for a minimum competency testing program, the nature of minimum competency tests, the test development methods, the assumptions of each equating model, time, and cost should be considered. Pre-equating should also be considered. Although it was not a goal of this study, the results of the study provided strong support for the methodology used in

pre-equating the forms of the AHSGE. The results of the study indicate that when the pre-equating is used, it may not be necessary to perform test score equating, except to test the pre-equating.

Factors Affecting Equating Results

Factors that may have affected the results of the equating include the set of common items, the estimation of the equipercentile scaled scores, the independence of test items, and the use of estimated true scores rather than estimated observed scores for the IRT models.

The set of common items for each subject area represented approximately one-half of the items comprising the test and approximately one-half of the competencies being tested. It has been recommended that the set of common items should be developed to represent as closely as possible a miniature version of the test. The sets of common items on the AHSGE were not developed with the purpose of being used for equating, but in order to create equivalent halves of the tests. The set of common items for each subject area test did correlate highly with the total test for each subject area, and it is unknown how the common items may have affected the results.

The results of the equipercentile model tend to be subjective because the final equated scores were matched using graphs and the judgement of the researcher. Due to the difficulty in estimating matching scores, they were only estimated to one-tenth of a point rather than one-hundredth of a point as in the other models. If the scaled scores for the other models were rounded to one-tenth, the results may be more comparable.

One of the assumptions of IRT is that the items on the test are independent of one another. This may not be true for all items on the AHSGE because in all subject areas some items use the same stimulus. It is not known if, or how, this may have affected the results of the equating. Another factor that may have affected the equating results is that true score rather than observed score estimates were used to transform IRT ability estimates. It could not be determined from previous research which method is more appropriate, and research is continuing in this area.

Recommendations for Further Research

Based on the results of this study, several recommendations should be considered. First, a replication of this study using two additional pre-equated forms of the AHSGE would determine if the results would be replicable. Second, a study using different sample sizes would provide information pertaining to the effect sample size has on the equating models. Third, this study did not examine clusters of items that use the same stem, even though this occurred in all three subject areas of the AHSGE. These items may not be independent and affect the results of the IRT equating. It may be useful to examine these item clusters and determine if they affect equating results. Fourth, after several administrations of the AHSGE, an earlier form is "rolled over" and used at a later administration. Sometimes items are substituted or the order of the items is changed. A study equating the original form and a rollover would provide information on whether rearrangement or substitution of items affects equating results.

References

- Alabama State Department of Education. (1988). Technical manual for the Alabama High School Graduation Examination. Tuscaloosa, AL: Evaluation and Assessment Laboratory, The University of Alabama.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.) Educational Measurement (pp. 508-600). Washington, DC: American Council on Education.
- Brennan, R. L. (1987). Introduction to problems, perspectives, and practical issues in equating. Applied Psychological Measurement, 11(3), 221-224.
- Cook, L. L., & Eignor, D. R. (1981, April). Score equating and item response theory: Some practical considerations. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.
- Cook, L. L., & Eignor, D. R. (1983, April). An investigation of the feasibility of applying item response theory to equate achievement tests. Paper presented at the annual meeting of the American Educational Research Association, Montreal.
- Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. Applied Psychological Measurement, 11(3), 225-244.
- Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart, and Winston.
- Cronbach, L. J. (1970). Essentials of Psychological Testing (3rd ed.). New York: Harper & Row.
- Cureton, E. E., & Tukey, J. W. (1951). Smoothing frequency distributions, equating tests, and preparing norms. (Abstract of presented paper.) American Psychologist, 6, 404.
- Eignor, D. R., Stocking, M. L., Cook, L. L. (1990). Simulation results of effects on linear and curvilinear observed- and true-score equating procedures of matching on a fallible criterion. Applied Measurement in Education, 3(1), 37-52.
- Fairbank, B. A. Jr. (1987). The use of presmoothing and postsmoothing to increase the precision of equipercentile equating. Applied Psychological Measurement, 11(3), 245-262.
- Gilmer, J. S. (1987, April). The effects of test disclosure on linear equating relationships under the common item nonequivalent groups design. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.
- Gulliksen, H. (1950.) Theory of mental tests. New York: Wiley.
- Hardy, R. A. (1984). Measuring instructional validity: A report of an instructional validity study for the Alabama High School Graduation Examination. Journal of Educational Measurement, 21(3), 291-301.

- Hills, J. R., Subhiyah, R. G., & Hirsch, T. M. (1988). Equating minimum-competency tests: Comparisons of methods. Journal of Educational Measurement, 25(3), 221-231.
- Holland, P. W., & Rubin, D. B. (1982). Introduction: Research on test equating sponsored by Educational Testing Service 1978-1980. In P. W. Holland and D. B. Rubin (Eds.) Test Equating (pp. 1-6). New York: Academic Press, Inc.
- Holmes, S. E. (1982). Unidimensionality and vertical equating with the Rasch model. Journal of Educational Measurement, 19, 139-147.
- Jaeger, R. M. (1981). Some exploratory indices for selection of a test equating method. Journal of Educational Measurement, 18(1), 23-38.
- Johnson, V. A. (1990). An analysis of methods used to reduce nonresponse bias in survey research. Unpublished doctoral dissertation, The University of Alabama, Alabama.
- Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. Psychometrika, 27, 59-72.
- Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. Journal of Educational Measurement, 18(1), 1-11.
- Lawrence, I. M., & Dorans, N. J. (1990). Effect on equating results of matching samples on an anchor test. Applied Measurement in Education, 3(1), 19-36.
- Livingston, S. A., Dorans, N. J., and Wright, N. K. (1990). What combination of sampling and equating methods works best? Applied Measurement in Education, 3(1), 73-96.
- Lockwood, R. E. (1985, April). The Alabama High School Graduation Examination experience: Technical concerns. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.
- Lockwood, R. E., & McLean, J. E. (1984, November). Equating of the Alabama High School Graduation Examination. In Technical aspects of implementing a high school graduation examination. Symposium conducted at the meeting of the Mid-South Educational Research Association, New Orleans.
- Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 39, 247-264.
- Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M. (1982). Item response theory and equating - a technical summary. In P. W. Holland and D. B. Rubin (Eds.) Test Equating (pp. 141-149). New York: Academic Press, Inc.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." Applied Psychological Measurement, 8(4), 453-461.

- Marco, G. L. (1981). Equating tests in an era of test disclosure. In B. F. Green (Ed.) New Directions for Testing and Measurement: Issues in Testing--Coaching, Disclosure, and Ethnic Bias (pp. 105-122). San Francisco, CA: Jossey-Bass.
- Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In D. Weiss (Ed.) New horizons in testing: Latent trait theory and computerized adaptive testing. New York: Academic Press.
- McLean, J. E., & Morse, B. (1985, April). Validating the Alabama High School Graduation Examination. In The Alabama High School Graduation Examination experience. Symposium conducted at the meeting of the National Council on Measurement in Education, Chicago.
- Modu, C. C. (1982, March). The robustness of latent trait model for achievement test score equating. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.
- Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. Journal of Educational Statistics, 8(2), 137-156.
- Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). In R. L. Linn (Ed.) Educational Measurement (3rd ed.) (pp. 221-262). New York: American Council on Education and Macmillan Publishing.
- Reinsch, C. H. (1967). Smoothing by spline functions. Numerische Mathematik, 10, 177-183.
- Schmitt, A. P., Cook, L. L., Dorans, N. J., and Eignor, D. R. (1990). Sensitivity of equating results to different sampling strategies. Applied Measurement in Education, 3(1), 53-72.
- Skaggs, G., & Lissitz, R. W. (1985, March). An exploration of the robustness of four test equating models. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois.
- Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. Review of Educational Research, 56(4), 495-529.
- SAS Institute Inc. (1985). SAS User's Guide: Statistics. Cary, NC: Author.
- Stocking, M. L., & Lord, F. M. (1982). Developing a common metric in item response theory. Research Report RR-82-5-ONR). Princeton, NJ: Educational Testing Service.
- Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.
- Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST 5, Version 1. Princeton, NJ: Educational Testing Service.